Yuan et al. BMC Genetics 2012, 13:83
http://www.biomedcentral.com/1471-2156/13/83

BMC Genetics

RESEARCH ARTICLE Open Access

Detection for gene-gene co-association via kernel canonical correlation analysis

Zhongshang Yuan[1†], Qingsong Gao[2†], Yungang He[3,4], Xiaoshuai Zhang[1], Fangyu Li[1], Jinghua Zhao[5] and Fuzhong Xue[1*]

Abstract

Background: Currently, most methods for detecting gene-gene interaction (GGI) in genomewide association studies (GWASs) are limited in their use of single nucleotide polymorphism (SNP) as the unit of association. One way to address this drawback is to consider higher level units such as genes or regions in the analysis. Earlier we proposed a statistic based on canonical correlations (CCU) as a gene-based method for detecting gene-gene co-association. However, it can only capture linear relationship and not nonlinear correlation between genes. We therefore proposed a counterpart (KCCU) based on kernel canonical correlation analysis (KCCA).

Results: Through simulation the KCCU statistic was shown to be a valid test and more powerful than CCU statistic with respect to sample size and interaction odds ratio. Analysis of data from regions involving three genes on rheumatoid arthritis (RA) from Genetic Analysis Workshop 16 (GAW16) indicated that only KCCU statistic was able to identify interactions reported earlier.

Conclusions: KCCU statistic is a valid and powerful gene-based method for detecting gene-gene co-association.

Keywords: Genome-wide association study (GWAS), Gene-gene co-association, Gene-gene interaction (GGI), Kernel canonical correlation analysis (KCCA)

Background

Genome-wide association studies (GWASs), which may involve a large number of single nucleotide polymorphisms (SNPs) on many individuals, are widely used to identify genetic variants underlying complex diseases or other types of traits. Although a primary interest in a GWAS is to identify SNPs associated with a trait of interest, it is important to consider the associate genes and their co-association as well. One form of co-association is epistasis, which was introduced approximately 100 years ago and generally defined as interactions among genes [1]. These are linked to gene-gene interactions (GGIs) which are often characterized to be functional, compositional and statistical [2]. The statistical definition was given by Fisher [3] and developed further by Cockerham [4] and Kempthorne [5], whereby the effect of GGIs is treated as deviation from additive genetic effects of single genes [6].

Methods to detect GGIs on the basis of the statistical definition include but are unlimited to logistic regression, multifactor dimensionality reduction [7], linkage disequilibrium (LD)-based [8,9] and entropy-based statistics [10,11], together with others implemented in PLINK[12], Tuning ReliefF [13], Random Jungle[14], BEAM[15] and BOOST[16]. However, most of these consider SNP as the unit of association, which has limitations and are insufficient for interpretation of GGI [17] which calls for consideration of higher level units such as genes or regions in the analysis. Gene-based analysis can account for multiple independent functional variants within genes with a potential increase of power to identify GGI. Earlier, Peng et al. [17] proposed a gene-based statistic (CCU statistic) for detecting gene-gene co-association based on canonical correlation analysis (CCA) in a case–control study, which was defined as joint effect of genes contributing to a binary trait and proved to have good performance on detecting gene-gene co-association or GGI. However, CCA can only

* Correspondence: xuefzh@sdu.edu.cn
†Equal contributors
[1]Department of Epidemiology and Health Statistics, School of Public Health, Shandong University, Jinan 250012, China
Full list of author information is available at the end of the article

detect linear correlation, and may be inappropriate for genomic data containing nonlinear structure. Recent years have witnessed considerable work and successes on kernel CCA (KCCA) as a nonlinear generalization of the classical CCA in machine learning, face recognition, data classification [18-20], and notably genomic data analysis by Yamanishi et al. [19]. We here construct a kernel CCU (KCCU) statistic for detecting gene-gene co-association and evaluate its performance via simulations and data analysis.

Methods

CCA

CCA is a classical multivariate method which concerns about linear dependencies between sets of variables. Let X_i, Y_i (i = 1, . . ., m) denote samples of measurements on m objects, We assume the data to be column centred. Let A be any $m \times n$ matrix then $L(A) = \{A\alpha | \alpha \in R^n\}$ will be referred to as the column-space and $L(A^T) = \{A^T\alpha | \alpha \in R^m\}$ the row-space of A. The aim of canonical correlation analysis is to determine vectors $v_j \in L(X^T)$ and $\omega_j \in L(Y^T)$ such that $a_j = Xv_j$ and $b_j = Y\omega_j$ are maximally correlated. $cor(a_j, b_j) = \frac{\langle a_j, b_j \rangle}{||a_j|| \bullet ||b_j||}$ with $\langle\rangle$ indicating inner product. Usually, this is formulated as a constrained optimization problem $\underset{v_j \in L(X^T), \omega_j \in L(Y^T)}{\arg\max} v_j{}^T X^T Y \omega_j$ subject to $v_j{}^T X^T X v_j = \omega_j{}^T Y^T Y \omega_j = 1$ which yields the first pair of canonical vectors (v_1, ω_1) and $a_1 = Xv_1$, $b_1 = Y\omega_1$ are the corresponding canonical variates and their correlation is called the maximum canonical coefficient. Pairs of canonical vectors (v_j, ω_j) can be recursively defined by maximizing similar expression and keeping subsequent variates orthogonal to those previously obtained. CCA can be interpreted as constructing pairs of factors from X and Y, respectively by linear combination of the variables involved, as a way to account for linear dependencies between sets of variables.

KCCA

KCCA generalizes CCA as follows: Objects x_i and y_i are first mapped to some Hilbert spaces H_x and H_y through mapping $\Phi_x(.)$ and $\Phi_y(.)$, CCA is then performed on images $\{\Phi_x(x_i)\}_{i=1}^m$ and $\{\Phi_y(y_i)\}_{i=1}^m$. Let K_x and K_y denote $m \times m$ kernel inner product matrices (also known as kernel gram matrices), constructed element-wise as $(K_x)_{ij} = \langle \Phi_x(x_i), \Phi_x(x_j) \rangle$ and $(K_y)_{ij} = \langle \Phi_y(y_i), \Phi_y(y_j) \rangle$. Analogous to CCA, the aim of KCCA is to find canonical vectors in terms of expansion coefficients $\alpha_j, \beta_j \in R^m$ as a constrained optimization problem $\underset{\alpha_j, \beta_j \in R^m}{\arg\max} \alpha_j{}^T K_X K_Y \beta_j$ subject to $\alpha_j{}^T K_X K_X \alpha_j = \beta_j{}^T K_Y K_Y \beta_j = 1$.

Explicit form for the mapping $\Phi_x(.)$ and $\Phi_y(.)$ are not always required but the kernel K_x and K_y need to be fixed. Common kernel functions include linear, polynomial, radial basis function (RBF), sigmoid [21], identical-by-state and weighted identical-by-state kernels [22]. It is worthwhile to note that these kernel functions generally have similar performance with parameters that are appropriately chosen.

Test statistic

Strategy analogous to CCU statistic was used to construct the KCCU statistic except that the maximum kernel canonical coefficient of the two genes, rather than the maximum canonical coefficient, was taken as a measure of gene-gene co-association in cases and controls. Let genotyped data of case–control study be $(X_1^D, X_2^D, \ldots, X_P^D)$ and $(Y_1^D, Y_2^D, \ldots, Y_q^D)$ for gene A and gene B for cases, and $(X_1^C, X_2^C, \ldots, X_P^C)$ and $(Y_1^C, Y_2^C, \ldots, Y_q^C)$ for controls. The maximum kernel canonical coefficient κr_D between $(X_1^D, X_2^D, \ldots, X_P^D)$ and $(Y_1^D, Y_2^D, \ldots, Y_q^D)$ obtained through KCCA could be considered as a measurement of gene-based gene–gene co-association in cases, and κr_C between $(X_1^C, X_2^C, \ldots, X_P^C)$ and $(Y_1^C, Y_2^C, \ldots, Y_q^C)$ be a measurement of gene–gene co-association in controls. The transformation analogous to Fisher's simple correlation coefficient transformation was done to κr_D and κr_C, i.e. $\kappa z_D = \frac{1}{2}(\log(1 + \kappa r_D) - \log(1 - \kappa r_D))$ and $\kappa z_C = \frac{1}{2}(\log(1 + \kappa r_C) - \log(1 - \kappa r_C))$.

The KCCU statistic for detecting statistical significance of the difference of gene-based gene-gene co-association between cases and controls can be defined as $U = \frac{\kappa z_D - \kappa z_C}{\sqrt{\mathrm{var}(\kappa z_D) + \mathrm{var}(\kappa z_C)}}$, which is approximately $N(0,1)$. With the difficulty in obtaining an explicit form for var (κz_D) and var(κz_C), a bootstrap procedure was employed. Seeing that the performance of kernel methods strongly relates to the choice of kernel functions and their parameters, we chose the RBF kernel owing to its flexibility in parameter specification [23]. In general, two approaches are popular: 1. via empirically assigning candidate values for the parameter(s) involved subject to a learning algorithm for the best performance; 2. via some cross-validation procedure. Both are computer intensive [24].

Data simulation

Simulation studies were conducted to assess the performance of KCCU relative to CCU under both the null (H_0) and alternative hypotheses (H_1), which were based on the HapMap data in the following steps:

Step 1. Phased haplotype (Phases 1 & 2 of CEU) data were downloaded from the HapMap web site (http://

snp.cshl.org) on two unlinked genome regions for generating the simulated genotypes. The *GNPDA2* region is on Chr 4: 44401210..44410098 involving six SNPs while *FAIM2* region is on Chr 12: 48571829..48583937 involving seven SNPs. Their LD patterns where shown in Figures 1 and 2 together with pairwise r^2.

Step 2. Based on data above, large samples with 100, 000 cases and 100, 000 controls were generated using software gs2.0 [25] under a two-locus interaction multiplicative effects model (see Additional file 1), treating the 2nd SNP of the first region and the SNP of the other as the causal variants and they were removed in the simulation to assess gene-gene co-association. The interaction odds ratio was set as 1.0 under H_0 and 1.1, 1.2, 1.3, 1.4, 1.5 under H_1. The SNPs in the regions were coded according to an additive genetic model. To further investigate the performance on causal SNPs with respect to minor allele frequency and LD, different SNP pairs from the two gene regions were defined as the casual variants.

Step 3. From the remaining SNPs, simulated data were sampled and CCU and KCCU performed under various sample sizes N ($N/2$ cases and $N/2$ controls, $N = 1000...5000$) with R package kernlab (http://cran.r-project.org/web/packages/kernlab/index.html). 500 simulations were repeated each with a significant level of 0.05.

Applications

The proposed KCCU statistic was applied to rheumatoid arthritis (RA) data from GAW16 Problem 1, consisting of 2,062 Illumina 550k SNP chips from 868 RA patients and 1,194 normal controls collected by the North American Rheumatoid Arthritis Consortium [26]. Three genes (*C5*, *ITGAV*, and *VEGFA*) on three different chromosomes were selected to detect gene-gene co-association in this work, involving eight, eight and four SNPs, respectively. Logistic regression test and the CCU statistic were also used. For each pair of genes, the statistic which yielded the minimum p value was recorded from all pairs of SNPs one on each gene. The significance of the statistic was compared to its empirical distribution generated from 1,000 permutations by permuting case–control labels [27] which is relatively easy compared to the "BY" method [28] for multiple testing adjustment.

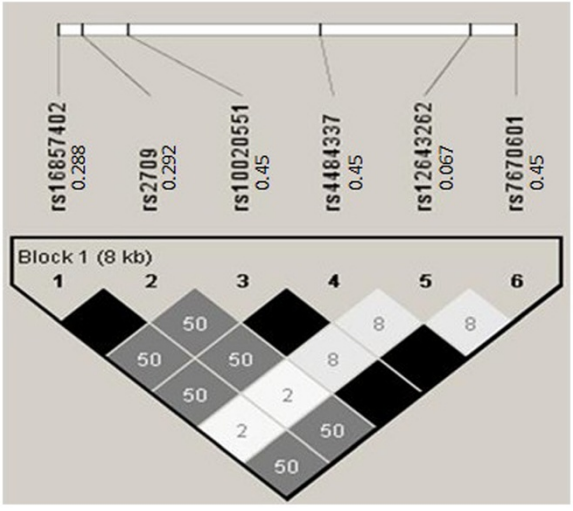

Figure 1 Pairwise r^2 among the six SNPs in the first region. The six SNPs are rs16857402, rs2709, rs10020551, rs4484337, rs12643262, and rs7670601. The values to the right of the 6 dbSNP IDs (rs# IDs) are the corresponding minor allele frequencies.

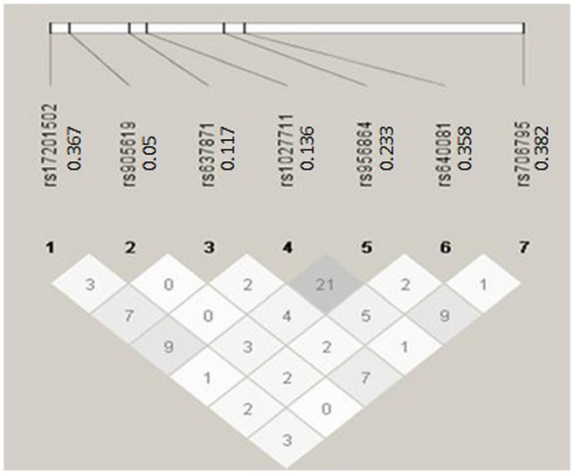

Figure 2 Pairwise r^2 among the seven SNPs in the first region. The seven SNPs are rs17201502, rs905619, r637871, rs1027711, rs956864, rs640081, and rs706795. The values to the right of the seven dbSNP IDs (rs# IDs) are the their minor allele frequencies.

Results

Simulation

Shown in Table 1 are simulation results under H_0. The KCCU statistic is normally distributed according to the one sample Kolmogorov-Smirnov test with the type I error rates of KCCU statistic being close to given nominal value ($\alpha = 0.05$) for different sample sizes. This indicates that the proposed statistic performs well under the null hypothesis.

Results on various interaction odds ratios and a sample size of 3,000 are shown in Figure 3, as with different sample sizes with an interaction odds ratio of 1.4 in Figure 4. It is clear that power of KCCU is a monotonically increasing function of sample size and interaction odds ratio. Figure 5 shows results with different SNP pairs defined as causal SNPs with an interaction odds ratio of 1.3. The power of KCCU statistic was higher than that of CCU statistic. Power as a function of interaction odds ratio for different sample size is provided as Additional file 1.

Table 1 Performance of CCU and KCCU under the null hypothesis

Sample size	CCU		KCCU	
	Type I Error	Normality Test (D)	Type I error	Normality Test (D)
1000	0.052	>0.55	0.049	>0.55
2000	0.051	>0.55	0.054	>0.55
3000	0.056	>0.55	0.052	>0.55
4000	0.048	>0.55	0.051	>0.55
5000	0.053	>0.55	0.050	>0.55

D, Kolmogorov-Smirnov D test.

Application

The performance of logistic regression test, CCU and KCCU statistics on pair-wise gene-gene co-association of three genes is shown in Table 2, which also contains results on the Gaussian RBF kernels with various parameter values (σ=0.05, 0.5, 5 and 50). Through KCCU the three genes were shown to have co-association with each other at significance level 0.05 regardless the parameter value, in contrast to the CCU statistics showing no significant co-association and none of the SNP pairs were significant under logistic regression test with correction for multiple testing.

Discussion

We have extended the CCU statistic to a new statistic KCCU, which can extract nonlinear correlation between two genes. Simulation studies show that both CCU and KCCU statistics performed well under null hypothesis with KCCU being more powerful than CCU with respect to significant level, sample size and relative risk. As results vary with user-defined kernel parameter, various parameters were used (the bandwidth parameter in RBF kernel) to RA data in GAW16 Problem 1, showing that the logistic regression test and CCU statistic failed to detect any interaction but KCCU statistics identified the pair-wise interactions among the three genes under various parameters. The interaction between *ITGAV* and *VEGF* genes has been identified by a rank method [29]. As suggested by a reviewer, it is critical to consider time-efficiency in genome-wide association studies to make the proposed methods practical. In our case, the computing time as required for KCCU was about 2.5 times slower than CCU, but nevertheless will still be

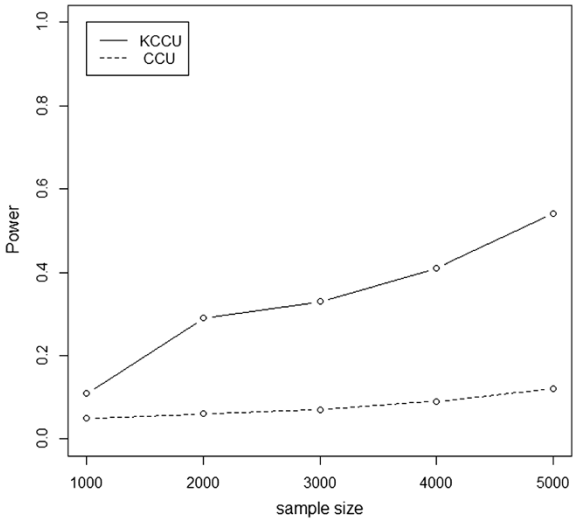

Figure 4 Power of CCU and KCCU statistics given an interaction odds ratio of 1.4 and different sample sizes.

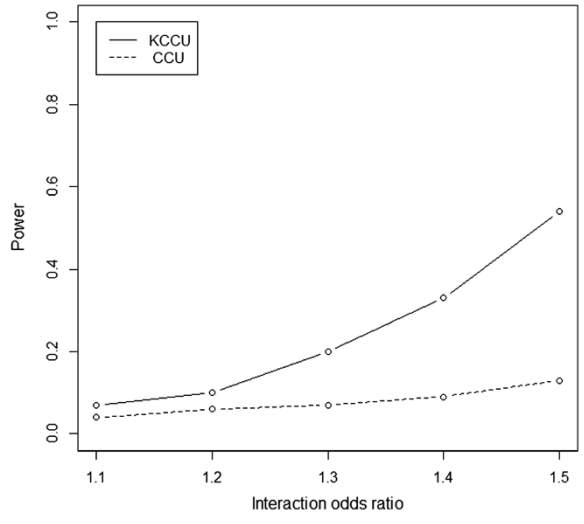

Figure 3 Power of CCU and KCCU statistics given different interaction odds ratios and a sample size of 3,000.

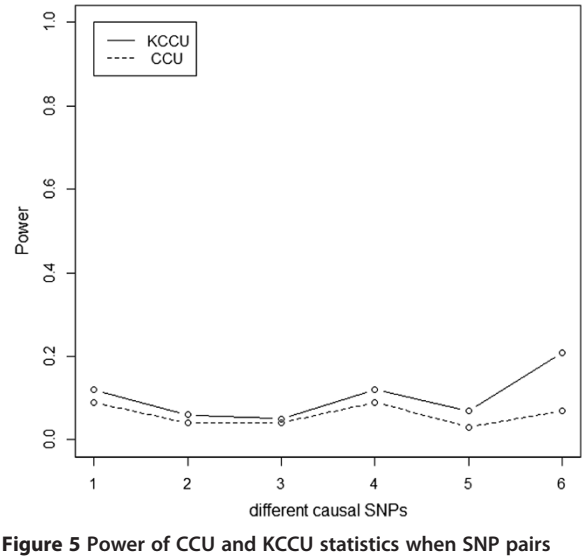

Figure 5 Power of CCU and KCCU statistics when SNP pairs from two regions are defined as casual variants at an interaction odds ratio of 1.3 and a sample size of 3,000.

Table 2 P-values of gene-gene co-association among *C5*, *ITGAV* and *VEGFA*

Co-association		*C5-ITGAV*	*C5-VEGFA*	*ITGAV-VEGFA*
Logistic regression		0.1015	0.1425	0.1840
CCU		0.5387	0.5325	0.8317
KCCU	σ=0.05	<0.001*	<0.001*	<0.001*
	σ=0.5	<0.001*	<0.001*	<0.001*
	σ=5	<0.001*	<0.001*	<0.001*
	σ=50	<0.001*	<0.001*	<0.001*

*significant at level 0.05.

feasible with the development as well as the extensive applications of multiprocessor and multithreading computational technique.

A reviewer has also suggested us to reiterate the relationship between gene-gene co-association and GGI which is readily available. GGI generally refers to the synergetic or antagonistic effect of two genes in addition to the summation of their independent effects on an outcome. To represent the interaction between two genes A and B in a case–control association study, a product term is customarily added to the logistic regression model $Logit(P) = \beta_0 + \beta_1 A + \beta_2 B + \gamma A \times B$ so that γ reflects both the direction and size of the interaction. This model implicitly assumes that gene A and gene B are independent so as to infer interaction (γ). However, it might well be that genes are correlated with each other in genetic networks to contribute to disease susceptibility, so the independence assumption is rarely ratified. Gene-gene co-association extends the concept of GGI in that it describes the generic joint distribution of two gene effects on disease or trait without assuming either independence or linear relationship. Here the measurement of the co-association between genes is based on the correlation between genes (such as CCU statistic and KCCU statistic), provides a measure of the contribution of two genes. As for two unlinked genes, their relationship can be described as either co-association or interaction. The reviewer has also brought to our attention to earlier work by Song and Nicolae [30] on imposing natural restrictions for the parameter space and discussion on the definition of "no interaction" between two unlinked loci as two loci being independent conditioned on the subject having the disease. In this paper, the null hypothesis of the proposed test is that there is no gene-gene co-association (i.e. GGI for two unlinked genes), the data under the null hypothesis are generated from the gs software with the interaction odds ratio parameter to be one.

Several issues remain to be resolved: the uncertainty to set the kernel function with appropriate parameters for each data, the undesirable performance of both CCU and KCCU with small interaction odds ratio (e.g. 1.1), and the possible failure of maximum kernel canonical correlation coefficient to represent gene-gene co-association.

Conclusions

KCCU statistic is a valid and powerful gene-based method for detecting gene-gene co-association compared to CCU and logistic regression test. Further work is needed to make its use in GWAS more practical.

Additional file

Additional file 1: Two-locus interaction multiplicative effects model.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZSY, QSG, YGH, XSZ, FYL, JHZ and FZX conceptualized the study, acquired and analyzed the data and prepared for the manuscript. All authors approved the final manuscript.

Acknowledgements

This work was supported by the grant from National Natural Science Foundation of China (30871392) and Young Talents Innovation Foundation of School of Public Health, Shandong University. We thank GAW16 and the North American Rheumatoid Arthritis Consortium for the RA data and two anonymous reviewers for suggestions which led to substantial improvement and clarification of the paper.

Author details

[1]Department of Epidemiology and Health Statistics, School of Public Health, Shandong University, Jinan 250012, China. [2]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, 13125 Berlin, Germany. [3]CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China. [4]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai 200031, China. [5]MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK.

Received: 18 March 2012 Accepted: 28 September 2012
Published: 8 October 2012

References

1. Bateson W: *Mendel's principles of heredity. 1909*, Mendel's principles of heredity. 902.
2. Phillips PC: **Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems.** *Nat Rev Genet* 2008, **9**:855–867.
3. Fisher RA: **The correlation between relatives on the supposition of Mendelian inheritance.** *Trans R Soc Edinb* 1918, **52**:399–433.
4. Cockerham CC: **An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present.** *Genetics* 1954, **39**(6):859.
5. Kempthorne O: **The correlation between relatives in a random mating population.** *Proc R Soc Lond B Biol Sci* 1954, **143**(910):103.
6. Cordell HJ: **Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans.** *Hum Mol Genet* 2002, **11**(20):2463.
7. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**(1):138–147.
8. Wu X, Jin L, Xiong M: **Composite measure of linkage disequilibrium for testing interaction between unlinked loci.** *Eur J Hum Genet* 2008, **16**(5):644–651.

9. Zhao J, Jin L, Xiong M: **Test for interaction between two unlinked loci.** *Am J Hum Genet* 2006, **79**(5):831–845.
10. Dong C, Chu X, Wang Y, Jin L, Shi T, Huang W, Li Y: **Exploration of gene-gene interaction effects using entropy-based methods.** *Eur J Hum Genet* 2007, **16**(2):229–235.
11. Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D: **An entropy-based approach for testing genetic epistasis underlying complex diseases.** *J Theor Biol* 2008, **250**(2):362–374.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: ***PLINK:*** **a tool set for wholegenome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
13. Moore J, White B: **Tuning reliefF for genomewide genetic analysis.** *Lect Notes Comput Sci* 2007, **4447:**166–175.
14. Schwarz D, Ko' Nig I, Ziegler A: **On safari to random jungle: A fast implementation of random forests for high dimensional data.** *Bioinformatics* 2010, **26:**1752–1758.
15. Zhang Y, Liu JS: **Bayesian inference of epistatic interactions in case–control studies.** *Nat Genet* 2007, **39:**1167–1173.
16. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang N, Yu W: **BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case–control Studies.** *Am J Hum Genet* 2010, **87:**325–340.
17. Peng Q, Zhao J, Xue F: **A gene-based method for detecting gene-gene co-association in a case–control association study.** *Eur J Hum Genet* 2009, **18**(5):582–587.
18. Zheng W, Zhou X, Zou C, Zhao L: **Facial expression recognition using kernel canonical correlation analysis (KCCA).** *Neural Networks IEEE Trans* 2006, **17**(1):233–238.
19. Yamanishi Y, Vert JP, Nakaya A, Kanehisa M: **Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis.** *Bioinformatics* 2003, **19**(suppl 1):i323.
20. Zepeda JAY, Davoine F, Charbit M: *Face tracking using canonical correlation analysis*, Visapp 2007: Proceedings of the Second International Conference on Computer Vision Theory and Applications, Volume Iu/Mtsv, Volume 559. 2007:396–402.
21. Liu Z, Chen D, Bensmail H: **Gene expression data classification with Kernel principal component analysis.** *J Biomed Biotechnol* 2005, **2:**155–159.
22. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: **Powerful SNP-set analysis for case–control genome-wide association studies.** *Am J Hum Genet* 2010, **86**(6):929–942.
23. Nguyen VH, Golinval JC: **Fault detection based on Kernel Principal Component Analysis.** *Eng Struct* 2010, **32**(11):3683–3691.
24. Zhang DQ, Zhou ZH, Chen SC: **Adaptive kernel principal component analysis with unsupervised learning of kernels**. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)*. 2006.
25. Li J, Chen Y: **Generating samples for association studies based on HapMap data.** *BMC Bioinforma* 2008, **9**(1):44.
26. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LRL, *et al:* **TRAF1-C5 as a risk locus for rheumatoid arthritis - A genomewide study.** *New Engl J Med* 2007, **357**(12):1199–1209.
27. Westfall PH, Young SS: *Resampling-based multiple testing*. New York: Wiley; 1993.
28. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Stat* 2001, **29:**1165–1188.
29. Huang CH, Cong L, Xie J, Qiao B, Lo SH, Zheng T: **Rheumatoid arthritis-associated gene-gene interaction network for rheumatoid arthritis candidate genes.** *BMC Proc* 2009, **3**(Suppl 7):S75.
30. Song M, Nicolae D: **Restricted parameter space models for testing gene-gene interaction.** *Genet Epidemiol* 2009, **33:**386–393.

doi:10.1186/1471-2156-13-83
Cite this article as: Yuan *et al.*: Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genetics* 2012 **13**:83.